**Course Seven**

# Google Advanced Data Analytics Capstone



# Instructions

Use this PACE strategy document to record your decisions and reflections as a data professional as you work through the capstone project. As a reminder, this document is a resource guide that you can reference in the future and a space to help guide your responses and reflections posed at various points throughout the project.

# Portfolio Project Recap

Many of the goals you accomplished in your individual course portfolio projects are incorporated into the Advanced Data Analytics capstone project including:

* Create a project proposal
* Demonstrate understanding of the form and function of Python
* Show how data professionals leverage Python to load, explore, extract, and organize information through custom functions
* Demonstrate understanding of how to organize and analyze a dataset to find the “story”
* Create a Jupyter notebook for exploratory data analysis (EDA)
* Create visualization(s) using Tableau
* Use Python to compute descriptive statistics and conduct a hypothesis test
* Build a multiple linear regression model with ANOVA testing
* Evaluate the model
* Demonstrate the ability to use a notebook environment to create a series of machine learning models on a dataset to solve a problem
* Articulate findings in an executive summary for external stakeholders

**Project proposal**

## **Salifort Motors Employee Turnover Prediction using Logistic Regression**

## **Overview**

Salifort Motors is seeking to understand the factors that cause employees to leave the company by analyzing internal employee data. This project will involve exploring this data, cleaning it, and constructing and evaluating a Logistic Regression model to predict the likelihood of an employee leaving. The ultimate goal is to provide data-driven insights and actionable recommendations to Salifort's HR department and leadership to help improve employee retention.

|  |  |  |
| --- | --- | --- |
| **Milestones** | **Tasks** | **PACE stages** |
| **1** | **Understand the business scenario & define the problem** | **PLAN** |
| **2** | **Data exploration & data cleaning (e.g., check duplicates, check outliers)** | **PLAN, ANALYZE** |
| **3** | **Determine appropriate model (Selecting Logistic Regression)** | **ANALYZE, CONSTRUCT** |
| **4** | **Construct the Logistic Regression model** | **CONSTRUCT** |
| **5** | **Check Logistic Regression assumptions (as applicable)** | **ANALYZE** |
| **6** | **Evaluate Logistic Regression model results using appropriate metrics** | **ANALYZE** |
| **7** | **Interpret results (coefficients) & share actionable steps** | **EXECUTE** |

**Data Project Questions & Considerations**

**PACE: Plan Stage**

**Foundations of data science**

* **Who is your audience for this project?**

The primary audience is the Salifort Motors Senior Leadership team and the Human Resources (HR) department

* **What are you trying to solve or accomplish? And, what do you anticipate the impact of this work will be on the larger business need?**

The project aims to address high employee turnover by analyzing HR data to understand its key drivers and building a Logistic Regression model to predict which employees are likely to leave. The anticipated impact is significant cost savings from reduced hiring and training, improved overall employee retention and satisfaction, and better workforce stability for Salifort Motors

* **What questions need to be asked or answered?**

1.What are the main factors contributing to an employee's decision to leave the company?

2. Can we reliably predict which employees are at high risk of turnover using the available data?

3. Based on the findings, what actionable recommendations can be provided to HR and leadership to

help retain employees?

* **What resources are required to complete this project?**

**Data**: The provided HR dataset (HR\_capstone\_dataset.csv) containing employee informations.

**Technology**: A Python environment (like the Jupyter Notebook provided in the lab) with standard data science libraries such as pandas, numpy, matplotlib, seaborn, and scikit-learn.

**Personnel:** Data analysis skills .

* **What are the deliverables that will need to be created over the course of this project?**

A completed PACE strategy document (this document).

A Jupyter Notebook containing all Python code for data cleaning, EDA, Logistic Regression model building, and evaluation.

An Executive Summary presenting the key findings, model performance (benefits and limitations), and actionable recommendations for the stakeholders.

**Get Started with Python**

* How can you best prepare to understand and organize the provided information?

The best way to prepare to understand and organize the provided data is by loading it into a structure

format like a pandas DataFrame, then using initial methods such as .head(), .info(), and .describe() to quickly inspect its structure, data types, summary statistics, and identify potential initial issues.

* What follow-along and self-review codebooks will help you perform this work?

The primary resource is typically the provided lab environment or Jupyter Notebook itself, which serves as a step-by-step guide and a space to write and review the code for data manipulation, analysis, and modeling. Referencing documentation for key libraries like pandas and scikit-learn also acts as a form of codebook and self-review resource.

* What are a couple additional activities a resourceful learner would perform before starting to code?

A resourceful learner might review the documentation for the main libraries they plan to use (e.g., pandas, scikit-learn) to refresh their knowledge of key functions, and search for examples of code for similar data analysis or machine learning tasks, such as employee turnover prediction, to get ideas for structuring their own code and visualizations

**Go Beyond the Numbers: Translate Data into Insights**

* What are the data columns and variables and which ones are most relevant to your deliverable?

The dataset contains 10 columns: satisfaction\_level, last\_evaluation, number\_project, average\_monthly\_hours, time\_spend\_company, Work\_accident, left, promotion\_last\_5years, Department, and salary. All are relevant, but the target variable left and those identified in EDA as strongly associated with turnover (like satisfaction\_level, average\_monthly\_hours, number\_project, tenure, promotion\_last\_5years, department, and salary) are most relevant to predicting turnover and providing actionable insights for the deliverable.

* What units are your variables in?

Variables have different units: satisfaction\_level and last\_evaluation are on a scale from 0 to 1. number\_project is a count. average\_monthly\_hours is hours per month. time\_spend\_company (renamed to tenure) is in years. Work\_accident, left, and promotion\_last\_5years are binary indicators (0 or 1). Department and salary are categorical labels.

* What are your initial presumptions about the data that can inform your EDA, knowing you will need to confirm or deny with your future findings?

Initial presumptions included that employees with lower satisfaction, higher workloads (hours, projects), longer tenure, lower salaries, and lack of promotion might be more likely to leave. We also presumed there would be some class imbalance in the turnover variable.

* Is there any missing or incomplete data?

Based on the df.info() and df.isna().sum() outputs, there is no missing data in this dataset.

* Are all pieces of this dataset in the same format?

The dataset contains a mix of numerical (int64, float64) and object (str) data types. Within each column, the data appears consistent, but categorical columns needed encoding for modeling.

* Which EDA practices will be required to begin this project?

Required EDA practices included loading the data, initial inspection (.head(), .info(), .describe()), checking for and handling duplicate rows, analyzing distributions using value counts and visualizations, and exploring relationships between variables through plots and correlation analysis.

**The Power of Statistics**

* What is the main purpose of this project?

The main purpose is to identify the reasons why employees leave Salifort Motors and to develop a predictive model that can forecast employee turnover to help improve retention.

* What is your research question for this project?

A central research question is: What factors significantly contribute to employee turnover, and can a model accurately predict which employees are at high risk of leaving the company?

* What is the importance of random sampling? In this case, what is an example of sampling bias that might occur if you didn’t use random sampling?

Random sampling is important to ensure the collected data is representative of the overall employee population at Salifort Motors, allowing the findings to be generalized. If random sampling was not used, a sampling bias could occur; for example, surveying only employees from specific departments might lead to conclusions about turnover drivers that do not apply to the entire company.

**Regression Analysis: Simplify Complex Data Relationships**

* Who are your stakeholders for this project?

The primary stakeholders are the Salifort Motors HR department and senior leadership. Future potential employers are also stakeholders.

* What are you trying to solve or accomplish?

The project aims to understand employee turnover drivers and build a model to predict which employees are likely to leave, to help improve retention

* What are your initial observations when you explore the data?

initial observations will involve checking the dataset's size, variable types, looking for missing values or duplicates, and getting a first sense of variable distributions and the target variable's balance.

* What resources do you find yourself using as you complete this stage? (Make sure to include the links.)

Resources will include the project description, data documentation, and guides on data analysis and modeling libraries as well as resource from kaggle link: https://www.kaggle.com/datasets/mfaisalqureshi/hr-analytics-and-job-prediction?select=HR\_comma\_sep.csv

* Do you have any ethical considerations in this stage?

Yes, ethical considerations involve ensuring data privacy and security, being mindful of potential biases, and planning for responsible use of insights.

**The Nuts and Bolts of Machine Learning**

* What am I trying to solve?

The task is to predict which employees are likely to leave the company, a binary classification problem.

* What resources do you find yourself using as you complete this stage?

Resources include documentation for machine learning libraries like Scikit-learn and potentially XGBoost, online tutorials for implementing classification models, and platforms like Stack Overflow for troubleshooting.

* Is my data reliable?

Based on initial checks, the data appears reasonably clean with no missing values and handle duplicates.However, as self-reported survey data, its inherent reliability might be less than objective performance metrics.

* Do you have any additional ethical considerations in this stage?

Additional ethical considerations involve ensuring fairness and avoiding bias in the predictions made by the machine learning model, and considering the explainability and responsible use of the model's outputs.

* What data do I need/would I like to see in a perfect world to answer this question?

In an ideal scenario, additional data such as detailed compensation history, manager feedback, team dynamics, training participation records, and objective performance metrics would provide a more comprehensive view to improve prediction accuracy.

* What data do I have/can I get?

The project provides a specific HR dataset with 10 columns of employee information, including satisfaction, evaluation, projects, hours, tenure, accident, left, promotion, department, and salary.

* What metric should I use to evaluate success of my business objective? Why?

For predicting employee turnover, key metrics are Precision, Recall, F1-score, and AUC, especially Recall and F1-score for the minority class (employees who leave). Recall is crucial because missing an employee who is leaving (False Negative) is often more costly to the business than a false alarm (False Positive), and F1-score balances precision and recall. AUC provides a robust measure of the model's overall ability to distinguish between leavers and stayers.

**Data Project Questions & Considerations**

**PACE: Analyze Stage**

**Get Started with Python**

* Will the available information be sufficient to achieve the goal based on your intuition and the analysis of the variables?

Yes, the available dataset containing employee information on satisfaction, performance evaluation, workload, tenure, promotions, department, and salary was sufficient to perform exploratory data analysis, identify key factors associated with employee turnover, and build a predictive model that showed some ability to distinguish between employees who stayed and those who left, thus allowing us to address the main goal of understanding and predicting turnover.

**Go Beyond the Numbers: Translate Data into Insights**

* What steps need to be taken to perform EDA in the most effective way to achieve the project goal?

Effective EDA involved loading and inspecting the data, handling data quality issues like duplicates and outliers, calculating descriptive statistics, and creating visualizations (histograms, box plots, scatter plots, heatmaps) to explore distributions and relationships between features and the target variable (turnover)

* Do you need to add more data using the EDA practice of joining? What type of structuring needs to be done to this dataset, such as filtering, sorting, etc.?

Based on the available data, joining with external datasets was not necessary to address the core project goal. Structuring included dropping duplicate rows, filtering out rows with outliers in the tenure variable for specific model preparation, and encoding categorical variables for modeling.

* What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?

Initial assumptions favored clear and interpretable visualizations like bar charts, histograms, box plots, and scatter plots to show distributions and relationships between variables, and heatmaps for correlations, as these are generally easy for stakeholders to understand

**The Power of Statistics**

* Why are descriptive statistics useful?

Descriptive statistics are useful because they provide a concise summary of the main features of a dataset, such as the central tendency (mean, median), variability (standard deviation, range), and distribution shape, which helps in understanding the data before deeper analysis or modeling. We used them to summarize numerical columns and understand proportions in binary variables.

* What is the difference between the null hypothesis and the alternative hypothesis?

The null hypothesis (H0​) is a statement of no effect or no relationship, often representing the status quo or a statement to be tested for possible rejection. The alternative hypothesis (H1​ or Ha​) is a statement that contradicts the null hypothesis, proposing that there is an effect or a relationship. In statistical testing, you gather evidence to see if there is enough support to reject the null hypothesis in favor of the alternative

**Regression Analysis: Simplify Complex Data Relationships**

* What are some purposes of EDA before constructing a multiple linear regression model?

The purposes of EDA before constructing a model like logistic regression were to understand the data's structure and variables, identify data quality issues like duplicates and outliers, discover patterns and relationships between features and the target variable through visualizations and statistics, and gain insights to inform feature selection and guide model choice.

* Do you have any ethical considerations in this stage?

Yes, ethical considerations in the Analyze stage involved handling employee data with privacy during exploration and cleaning, being aware of potential biases reflected in the data, and making transparent decisions regarding data cleaning and outlier handling that could impact subsequent analysis.

**The Nuts and Bolts of Machine Learning**

* What am I trying to solve? Does it still work? Does the plan need revising?

We are trying to predict employee turnover. Based on EDA, the data appears relevant and the plan to build a predictive model still works, although initial model performance might necessitate revisiting some data preparation or model selection decisions, potentially revising the plan for improvement

* Does the data break the assumptions of the model? Is that ok, or unacceptable?

For Logistic Regression, the data does break some assumptions like multicollinearity and outliers. This is not ideal, but depending on the severity and impact on performance, it might be acceptable for a baseline, though ideally, steps would be taken to mitigate this or use models less sensitive to these issues.

* Why did you select the X variables you did?

The independent variables (X variables) were selected because EDA revealed their relationships with the target variable (employee turnover), indicating they are likely factors that influence whether an employee leaves the company.

* What are some purposes of EDA before constructing a model?

The purposes of EDA were to understand the data, identify quality issues, discover patterns and relationships between features and the target variable, and gain insights to inform feature selection and guide model choice.

* What has the EDA told you?

EDA has revealed key turnover drivers, data quality issues like duplicates and outliers, the need for categorical encoding, and patterns in variable distributions and relationships.

* What resources do you find yourself using as you complete this stage?

Resources included documentation for data manipulation and visualization libraries (Pandas, Matplotlib, Seaborn) and potentially machine learning libraries (Scikit-learn) for initial data prep functions.

* Do you have any ethical considerations in this stage

Ethical considerations involved handling employee data with privacy during exploration and cleaning, being aware of potential biases reflected in the data, and making transparent decisions during data cleaning processes.

**Data**

**Project Questions & Considerations**

**PACE: Construct Stage**

**Get Started with Python**

* Do any data variables averages look unusual?

When examining descriptive statistics, no variable averages appeared extremely unusual in a way that suggested data errors, although the average monthly hours did seem high compared to a standard work month, which aligns with the potential overwork issue identified in EDA

* How many vendors, organizations or groupings are included in this total data?

Based on the dataset, there is one organization (Salifort Motors). Key groupings within the data include the different departments (Sales, Technical, Support, etc.) and salary levels (low, medium, high)

**Go Beyond the Numbers: Translate Data into Insights**

* What data visualizations, machine learning algorithms, or other data outputs will need to be built in order to complete the project goals?

The primary machine learning algorithm built was Logistic Regression. Data outputs included the confusion matrix and calculation of various performance metrics like precision, recall, F1-score, and AUC.

* What processes need to be performed in order to build the necessary data visualizations?

Building necessary data visualizations involved using Python libraries like Matplotlib and Seaborn, selecting the appropriate plot type for the variables being examined (e.g., scatterplot for two numerical variables, histogram for distributions, boxplot for distributions across categories, heatmap for correlations), mapping the relevant variables to the plot's axes and visual properties (like color based on the target variable), adding titles and labels, and displaying the plot.

* Which variables are most applicable for the visualizations in this data project?

Variables most applicable for visualizations included satisfaction\_level, last\_evaluation, number\_project, average\_monthly\_hours, tenure, work\_accident, promotion\_last\_5years, department, and salary. These were frequently visualized individually to understand distributions or in relation to the target variable left to explore potential drivers of turnover.

* Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

Based on the analysis in the Analyze stage, the dataset was found to have no missing data, so the plan for dealing with missing data became to simply confirm its absence rather than implementing imputation or removal strategies.

**The Power of Statistics**

* How did you formulate your null hypothesis and alternative hypothesis?

While formal hypothesis testing was not the primary method in this predictive modeling project, the underlying concept involves considering a null hypothesis where there is no relationship between the independent variables and employee turnover, versus an alternative hypothesis where such a relationship exists. In predictive modeling, the focus shifts to how well the model can capture these relationships to make accurate predictions rather than formally testing the hypotheses.

* What conclusion can be drawn from the hypothesis test?

As a formal hypothesis test was not conducted as part of the model building process in this lab, a conclusion from such a test cannot be drawn. Instead, conclusions about the importance of variables and their relationship to turnover are drawn from exploratory data analysis insights and model performance metrics.

**Regression Analysis: Simplify Complex Data Relationships**

* Do you notice anything odd?

During data preparation for Logistic Regression, addressing categorical encoding and potential outliers in tenure required specific steps. In the model itself, the lower precision and recall for predicting the minority class were initially notable compared to overall accuracy.

* Can you improve it? Is there anything you would change about the model?

Yes, the Logistic Regression model can likely be improved, particularly in its ability to identify leavers. Changes would involve potentially trying different models like Random Forest or XGBoost, hyperparameter tuning, and addressing class imbalance or assumptions violations more directly.

**The Nuts and Bolts of Machine Learning**

* Is there a problem? Can it be fixed? If so, how?

The problem is the model's limited effectiveness in predicting the minority class (leavers) based on initial evaluation metrics. Yes, it can likely be improved. This can be addressed by trying alternative algorithms like Random Forest or XGBoost, which may capture complex patterns better, optimizing model settings through hyperparameter tuning, or using techniques to address the class imbalance directly.

* Which independent variables did you choose for the model, and why?

All variables in the prepared data except the target left were chosen as independent variables because EDA indicated their relationship with employee turnover, and they were encoded/prepared to be suitable for the Logistic Regression model.

* How well does your model fit the data? (What is my model’s validation score?)

The Logistic Regression model shows a reasonable overall fit with an AUC score of 0.8403, which is a key validation score. However, its fit for predicting the minority class is limited, as indicated by lower precision (0.44) and recall (0.26) for employees who leave.

* Can you improve it? Is there anything you would change about the model?

Yes, the Logistic Regression model can likely be improved. To improve its ability to identify leavers, one might try different classification models (e.g., Random Forest, XGBoost), conduct hyperparameter tuning, or address class imbalance techniques.

* Do you have any ethical considerations in this stage?

Yes, ethical considerations in the Construct stage involved being mindful of potential biases in the data that the model could learn and ensuring the fairness of the model's predictions, particularly across different employee groups.

**Data Project Questions & Considerations**

**PACE: Execute Stage**

**Get Started with Python**

* Given your current knowledge of the data, what would you initially recommend to your manager to investigate further prior to performing an exploratory data analysis?

Given the likely impact of employee satisfaction on turnover (as hinted by basic descriptive statistics even before deep analysis), an initial recommendation might be to investigate current employee satisfaction levels more thoroughly, perhaps through targeted surveys or discussions in departments suspected of having high turnover.

* What data initially presents as containing anomalies?

Based on initial data checks, a significant number of duplicate rows were identified as anomalies that required handling. Outliers in the tenure variable were also identified as data points that warrant attention or specific handling depending on the modeling approach.

* What additional types of data could strengthen this dataset?

Additional data that could strengthen the dataset and potentially improve prediction accuracy includes more detailed compensation history, information on manager effectiveness, participation in training and development programs, specific career path progression within the company, and possibly more granular data on workload beyond just hours and project counts

**Go Beyond the Numbers: Translate Data into Insights**

* What key insights emerged from your EDA and visualizations(s)?

EDA and visualizations revealed that low satisfaction levels, high average monthly hours, and having either a low (2) or very high (6-7) number of projects are strongly associated with employee turnover. Additionally, visualizations highlighted the high turnover rate at tenure years 6+ and suggested a relationship between longer tenure, lower salary brackets, and leaving. The impact of promotion on retention and turnover volume by department were also key visual insights.

* What business recommendations do you propose based on the visualization(s) built?

Based on the visualizations, business recommendations include addressing low satisfaction through targeted efforts, managing employee workload to prevent burnout (especially for those with high projects/hours), investigating reasons for departure among employees with low project counts, and reviewing salary progression for long-tenured employees. Focusing retention efforts on departments with high turnover volume (Sales, Technical, Support) is also recommended.

* Given what you know about the data and the visualizations you were using, what other questions could you research for the team?

Other questions to research include a deeper investigation into the specific reasons behind low satisfaction and high hours/projects in certain groups, exploring factors not in the dataset like manager quality or training impact, or conducting a cost-benefit analysis of implementing specific retention strategies based on the findings.

* How might you share these visualizations with different audiences?

Visualizations could be shared with different audiences by tailoring the complexity and level of detail; for senior leadership, focus on high-level trends and key takeaways using clear charts in a concise report or presentation, while for HR analysts, provide more detailed plots and interactive visualizations in a notebook or dashboard that allows for deeper exploration.

**The Power of Statistics**

* What key business insight(s) emerged from your A/B test?

This project focused on building a predictive model for employee turnover rather than conducting an A/B test, so insights from an A/B test are not applicable in this context.

* What business recommendations do you propose based on your results?

Business recommendations based on the analysis and model results include addressing low employee satisfaction and high workloads to prevent burnout, reviewing salary and promotion structures for long-tenured employees, and focusing targeted retention efforts on high-turnover departments like Sales, Technical, and Support.

**Regression Analysis: Simplify Complex Data Relationships**

* To interpret model results, why is it important to interpret the beta coefficients?

In Logistic Regression, interpreting the beta coefficients is important because they represent the change in the log-odds of the outcome variable (employee leaving) for a one-unit increase in the independent variable, holding others constant. This helps in understanding the direction and strength of the linear relationship between each feature and the likelihood of turnover.

* What potential recommendations would you make to your manager/company?

Potential recommendations include encouraging investment in HR analytics, implementing targeted retention programs based on identified risk factors, and reviewing company policies related to workload, compensation, and promotions to address key turnover drivers.

* Do you think your model could be improved? Why or why not? How?

Yes, the Logistic Regression model could likely be improved because its performance in predicting actual leavers (low recall) is limited. Improvement could involve trying other model types like Random Forest or XGBoost, optimizing hyperparameters, or addressing class imbalance techniques.

* What business recommendations do you propose based on the models built?

Business recommendations include addressing low employee satisfaction and high workloads to prevent burnout, reviewing salary and promotion structures for long-tenured employees, and focusing targeted retention efforts on high-turnover departments like Sales, Technical, and Support.

* What key insights emerged from your model(s)?

Key insights confirmed that low satisfaction, high workload/projects, longer tenure, and lack of promotion are significant turnover drivers. The model evaluation highlighted the model's overall ability to distinguish classes (AUC 0.84) but also its limitations in precisely identifying leavers (low precision/recall).

* Do you have any ethical considerations at this stage?

Yes, ethical considerations in the Execute stage involve interpreting results responsibly, avoiding discriminatory actions based on predictions, ensuring transparency about model limitations with stakeholders, and protecting employee data privacy.

**The Nuts and Bolts of Machine Learning**

* What key insights emerged from your model(s)?

Key insights from the Logistic Regression model and prior EDA confirmed that factors like low satisfaction, high workload (hours/projects), longer tenure, and lack of promotion are significant predictors of employee turnover. Model evaluation highlighted its overall ability to distinguish leavers from stayers (AUC 0.84), but also its limitations in precisely identifying true leavers (low recall 0.26) compared to false alarms (precision 0.44).

* What are the criteria for model selection?

Criteria for model selection include the type of prediction task (binary classification), the characteristics of the data (mix of numerical/categorical, imbalance, non-linearity, multicollinearity), the desired balance between interpretability and predictive performance, and the specific business needs (e.g., prioritizing minimizing false negatives).

* Does my model make sense? Are my final results acceptable?

The Logistic Regression model makes sense in that its performance metrics (like AUC) show it's better than random chance, and its insights align with expected turnover drivers. However, whether the final results are "acceptable" depends on the specific business requirements for precision and recall, especially for the minority class; given the low recall (26%), it might not be acceptable if the priority is to identify most at-risk employees.

* Were there any features that were not important at all? What if you take them out?

Based on the Logistic Regression model (which gives some indication via coefficient magnitude, though not formal feature importance), and correlation analysis, some features showed weaker linear relationships with turnover than others. Taking out features with low importance might simplify the model or address multicollinearity, but it could also reduce performance if the feature has non-linear relationships or interactions not captured by the simple linear model

* Given what you know about the data and the models you were using, what other questions could you address for the team?

Other questions include investigating root causes of turnover within specific departments or roles, exploring the impact of factors not in the dataset (e.g., manager quality, training), assessing the financial impact of turnover, or developing models to predict when employees might leave.

* What resources do you find yourself using as you complete this stage?

Resources included documentation for evaluation metrics in Scikit-learn, visualization tools for presenting results, guides on interpreting model outputs and performance metrics, and resources on ethical considerations in machine learning.

* Is my model ethical?

Whether the model is considered ethical depends on its development and use. Building the model involved considering potential biases in the data and aiming for fair predictions. Ethical use requires transparent communication of its limitations and potential errors, avoiding discriminatory applications of its predictions, and ensuring human oversight in any decisions informed by the model

* When my model makes a mistake, what is happening? How does that translate to my use case?

When the model makes a mistake, it's incorrectly classifying an employee as either staying when they leave (False Negative) or leaving when they stay (False Positive). In this use case, a False Negative means missing an employee who is about to leave (a missed opportunity for intervention), while a False Positive means incorrectly flagging a stable employee as a risk (potentially wasted resources or unnecessary concern).